Glossary and Metadata Tags: The Internet Archive—Beyond the Wayback Machine

Core Organization & Concepts

Internet Archive: 501(c)(3) nonprofit research library founded in 1996 by Brewster Kahle; provides universal access to digitized materials, web archives, and knowledge resources. Mission: "Universal Access to All Knowledge."

Wayback Machine: Web archiving service that captures and preserves snapshots of websites across time; preserves 866+ billion web pages but represents only one of many Internet Archive services.

Digital Preservation: Systematic process of maintaining and ensuring long-term accessibility of digital materials, including web content, software, media, and documents.

Brewster Kahle: Founder of Internet Archive and pioneering digital librarian; visionary behind "digital Library of Alexandria."

Collection Types and Services

Open Library: Sister project cataloging 3+ million books with open, editable library catalog; facilitates controlled digital lending between libraries.

Internet Archive Scholar: Fulltext search index and discovery service providing access to 35+ million research articles, scholarly documents, and archived academic journals.

Live Music Archive: Specialized collection preserving 250,000+ concert recordings from 8,000+ artists spanning 20+ years; includes lossless and MP3-encoded formats.

TV News Archive: Global broadcast preservation service capturing 100+ television channels in 50+ countries and territories; archives 8 million broadcasts with searchable closed captions.

Archive-It: Institutional web archiving service (subscription-based) enabling libraries, museums, universities, and cultural organizations to preserve specialized website collections.

Community Webs: Archive-It program funded by Institute of Museum and Library Services enabling public libraries to document local history and community-oriented web content.

Emularity: Advanced software preservation technology enabling direct browser-based access to 250,000+ historical software programs without downloads or installation.

DAISY Format: Digital Accessible Information System standard for creating accessible ebooks and audiobooks for print-disabled and visually impaired users; maintained for 2+ million titles in Internet Archive.

Specialized Programs and Initiatives

Controlled Digital Lending (CDL): Library lending model enabling libraries to lend digital copies of books in their physical collections; central to Open Library lending but limited by 2023 litigation.

Print Disability Access: Internet Archive program providing specialized accessible formats and services for blind, visually impaired, and print-disabled users; coordinates with National Library Service for the Blind and Physically Handicapped.

Federal Depository Library Program: Network of 1,100+ libraries designated to preserve and provide public access to government documents; Internet Archive gained status July 2025.

Federal Depository Library Status: July 2025 designation of Internet Archive as official member of Federal Depository Library Program; recognizes digital-first approach as essential for government document preservation.

Decentralized Web Initiative: Internet Archive program promoting distributed web infrastructure and protocol development for resilient, censorship-resistant digital systems.

Media Formats and Collections

Web Archives: Preserved snapshots and versions of websites; 866+ billion unique pages captured.

Books and Texts: 47+ million items including public domain works, copyrighted materials through CDL, and digitized historical texts.

Audio Collections: 14+ million items including music concerts, interviews, field recordings, and speech content.

Video Collections: 13+ million items including feature films, newsreels, educational materials, and user-contributed content.

Television News Reports: 3+ million broadcasts with searchable closed captions; sourced from 100+ monitored channels globally.

Images: 5+ million items including Cover Art Archive, NASA imagery, museum photographs, and crowdsourced collections.

Software Programs: 1.2+ million items including MS-DOS games, Apple II software, obsolete operating systems, and HyperCard applications; accessible through Emularity.

Government Documents: Federal publications and materials; preservation enhanced by 2025 federal depository status.

Infrastructure and Operations

Scanning Centers: Distributed facilities globally processing 4,300+ books daily for digitization.

Data Storage: 100+ petabytes of stored digital materials maintained across infrastructure.

Staff: Approximately 140 full-time employees operating from San Francisco headquarters.

Preservation Technology: Includes web crawling, OCR (optical character recognition), format migration, emulation, and metadata standards.

Fatcat: Open catalog system maintaining bulk metadata with API access for scholarly and institutional integration.

Legal and Regulatory Concepts

Copyright Litigation: 2023 federal court ruling against controlled digital lending; forced removal of 500,000 books from lending program; appeals ongoing.

Marrakesh Treaty: International agreement establishing frameworks for cross-border access to materials in accessible formats; relevant to Internet Archive's print disability work.

DDoS Attacks: October 2024 distributed denial-of-service attacks rendered Internet Archive temporarily unavailable; highlighted infrastructure vulnerability.

Publisher Partnerships: Strategic agreements enabling licensed ebook lending; response to CDL litigation limiting unlicensed lending.

Funding and Sustainability

Nonprofit Status: 501(c)(3) classification enabling tax-deductible donations and grant funding.

Individual Donations: Primary funding mechanism from users and supporters.

Foundation Grants: Support from organizations including Alfred P. Sloan Foundation and competitive processes.

MacArthur Foundation 100&Change: Kahle's digitization proposal achieved semifinalist status; represents competitive funding model.

Alfred P. Sloan Foundation: Awarded $1 million grant to Internet Archive for digitization and preservation work.

Historical and Contextual References

"Digital Library of Alexandria": Founder Kahle's vision for Internet Archive; references ancient Library of Alexandria with metaphorical improvement ("better fire protection").

Universal Access to All Knowledge: Internet Archive's foundational mission statement.

Scholarly Record Preservation: Critical function addressing disappearance of research materials when publishers cease operations or institutional servers fail.

Government Transparency: Role in ensuring perpetual public access to federal publications and documentation.

Cultural Heritage Preservation: Archiving broadcast media, music performances, and historical software.

Digital Accessibility: Ensuring inclusive access for print-disabled and visually impaired populations.

Key Statistics and Measurements

80+ Million Objects: Total items preserved across all collections.

1+ Trillion Web Pages: Archived through Wayback Machine.

3+ Million Books: Open Library catalog size.

35+ Million Research Articles: Internet Archive Scholar index coverage.

250,000+ Concert Recordings: Live Music Archive collection.

100+ Television Channels: Monitored globally for TV News Archive.

5.65 Million Hours: Total broadcast content in TV News Archive.

250,000+ Software Programs: Accessible through Emularity.

1.2+ Million Accessible Titles: Available in DAISY format for print-disabled users.

4,300 Books Daily: Processing rate through distributed scanning centers.

100+ Petabytes: Data storage capacity.

140 Full-time Staff: Organizational employment.

Suggested Internet Archive Metadata Tags

For optimal discoverability and categorization, recommended tags include:

Primary Tags: digital-preservation | internet-archive | nonprofit-library | wayback-machine | web-archiving | digital-library

Collection Tags: open-library | internet-archive-scholar | live-music-archive | tv-news-archive | archive-it | software-preservation | emulation

Subject Tags: scholarly-communication | government-documents | digital-accessibility | print-disability | cultural-heritage | broadcast-media | music-preservation

Format Tags: web-archives | ebooks | audio | video | government-publications | historical-software | archival-metadata

Organizational Tags: brewster-kahle | federal-depository-library | nonprofit-infrastructure | digital-stewardship | knowledge-preservation

Functional Tags: controlled-digital-lending | community-web-archiving | scholarly-indexing | digital-preservation-technology | emulation-technology

Geographic/Scope Tags: global-coverage | broadcast-journalism | international-standards | united-states-government-documents

Legal/Policy Tags: copyright-litigation | digital-rights | marrakesh-treaty | print-disability-access | federal-depository-status